УДК 004.93'1

Н.Н. Масалитина
Учреждение образования «Гомельский государственный технический университет имени П.О. Сухого», г. Гомель, Республика Беларусь
Беларусь, 246746, г. Гомель, пр. Октября, 48, *masalitina@rambler.ru*

Классификация без учителя на основе неколичественно заданного критерия разделения классов

N.N. Masalitina
P. Sukhoi Gomel State Technical University, Gomel
Belarus Republic, 246746, Gomel, Pr. Octiabria, № 48

The Categorization Method without Learning Based on Qualitative Class Division Criteria

Н.Н. Масалітіна
Установа освіти «Гомельський державний технічний університет ім. П.О. Сухого», м. Гомель, Республіка Білорусь
Білорусь, 246746, м. Гомель, пр. Жовтня,48

Класифікація без вчителя на основі некількісних заданому критерію поділу класів

Предложен метод мультиклассовой классификации без учителя, основанный на аппарате алгебры логики. Метод предполагает исследование пространства возможных состояний объектов классификации на непротиворечивость, что позволяет существенно сократить размерность области существования объекта. В результате становится возможным применение качественного критерия разделения классов. Метод ориентирован на классификацию объектов, обладающих иерархической структурой управления.
Ключевые слова: мультиклассовая классификация, алгебра логики, качественные критерии

The multiclass categorization method without learning was proposed. Principles of Boolean algebra are in the base of the method. The method performs inconsistent descriptors combinations excluding. It allows reducing a dimensionality of an object existence area. As a result, the method allows using qualitative class division criteria. The method is orientated to the objects with a hierarchy structure.
Key Words: multiclass categorization, Boolean algebra, qualitative criteria.

Запропоновано метод мультикласової класифікації без вчителя, заснований на апараті алгебри логіки. Метод передбачає дослідження простору можливих станів об’єктів класифікації на несуперечність, що дозволяє істотно скоротити розмірність області існування об’єкта. У результаті стає можливим застосування якісного критерію поділу класів. Метод орієнтований на класифікацію об’єктів, що володіють ієрархічною структурою управління.
Ключові слова: мультикласова класифікація, алгебра логіки, якісні критерії.

Введение

Принятие решения во многих областях деятельности человека приводит к необходимости решения задачи классификации (распознавания образов): выявления принадлежности некоторого объекта к одной из групп, отличающихся по известному признаку.

Определение принадлежности объекта к некоторой группе требует анализа большого объема информации, часто приводит к сопоставлению противоречивых сведений, осложняется невозможностью проведения натурных экспериментов над рядом объектов управления. При этом в таких областях, как медицинская диагностика, идентификация личности, защита от атак компьютерных вирусов, управление производственными системами возникают особые требования к скорости распознавания, так как время на принятие решения и его практическое воплощение ограничено. Перечисленные особенности определяют необходимость математического моделирования и автоматизации процесса классификации объектов управления различной природы.

Существующие в настоящее время подходы к решению задач классификации можно выделить пары принципиально различных групп:

1) по количеству выделяемых классов: методы бинарной и мультиклассовой классификации;

2) по способу формализации классификатора: классификация с учителем и без учителя.

Бинарная классификация позволяет разделять множество исследуемых объектов на два подмножества (класса) [1], [2]. Задачи, требующие выделения более чем двух подмножеств, решаются на основе методов мультиклассовой классификации [3], [4].

Методы классификации с учителем реализуют формализацию правила разделения подмножеств на основе обобщения информации о значениях ряда показателей (дескрипторов) объектов, принадлежность которых к данным подмножествам задана [1], [3], [4]. При этом необходима достоверная информация о составе классов и о значениях всех входных и выходных параметрах модели по достаточно большому множеству объектов, которое принято называть обучающей выборкой или группой эталонов. Такие методы допускают использование описаний объектов классификации и правил их разделения на группы в любой форме: количественной и качественной, непрерывной и дискретной. Основная сложность при использовании методов, основанных на обучении, определяется сложностью обеспечения репрезентативности обучающей выборки в особенности в долгосрочной перспективе.

Методы второй группы позволяют выявлять среди исследуемых объектов наиболее похожие по некоторым показателям. В этом случае классификация выполняется на основе вычисления различных расстояний между значениями дескрипторов [5], [6]. Природа родства между объектами различных классов остается необъяснимой. Состав классов может быть и неизвестен на начальных этапах исследования, однако критерии их разделения обязательно должны быть заданы в количественной форме. Преимуществом методов классификации без учителя является скорость и независимость от истории исследуемых объектов.

Практика управления часто приводит к задачам классификации объектов, для которых характерна высокая скорость изменения структуры, поэтому неэффективны методы, требующие обучения. Необходимость классификации по критерию, заданному в неколичественной форме, либо невозможность заранее сформировать полный перечень классов приводит к задаче, решение которой на основе существующих методов неэффективно.

Задача классификации с неизвестным составом алфавита классов и качественным критерием их разделения решена автором на примере управления устойчивостью систем, обладающих иерархической структурой. К числу таких систем относятся субъекты хозяйствования, компьютерные сети, функциональные системы живых организмов в аспекте управления ходом лечения заболеваний и проч. В качестве

неколичественного критерия разделения классов выбрано следующее условие: элементы различных классов должны принципиально отличаться возможностями и целями управления устойчивостью.

Целю работы – получить метод мультиклассовой классификации систем с иерархической структурой управления (СИСУ), допускающий использование неколичественного описания критериев разделения классов и не требующий априорной информации о составе множества классов и о принадлежности к классам ряда реальных объектов.

Постановка задачи – формализовать правило, разделяющее множество СИСУ на n классов b_i $(i=1..n,$ где n – количество классов), таких, что элементы каждого класса b_i отличались от элементов каждого другого класса b_j $(j=1..n,\ j \neq i)$ по критерию d_{ij} – в состав множества Mt_i управляющих воздействий, необходимых для повышения устойчивости объекта класса b_i, входят элементы, не входящие в множество Mt_j управляющих воздействий, необходимых для элементов класса b_j.

Методы исследования

Выбор методов исследования определяется спецификой исследуемой предметной области – управления устойчивостью систем, характеризующихся иерархическими взаимосвязями. Данное исследование строится на информации о исправности (надлежащем функционировании) отдельных элементов системы и о используемых при управлении механизмах. Особенностью указанных показателей является удобство представления в двоичной форме (1 – элемент исправен, управляющее воздействие применяется, механизм достаточно эффективен, 0 – неисправен, не используется, не эффективен) и как следствие возможность применения методов алгебры логики. Иерархические взаимосвязи также легко представляются в виде матриц инцидентности, представляемых в двоичной форме.

Двоичная форма показателей, используемых для описания состояния системы, определяет выбор методов алгебры логики. К числу существенных преимуществ данных методов относится легкая автоматизация и аппаратная реализация, хорошая проработанность приемов доказательства и опровержения [2].

Результаты исследования

Разработанный автором метод мультиклассовой классификации предполагает последовательное прохождение ряда основных этапов, представленных на рис. 1.

Первый этап метода предполагает составление описания СИСУ с помощью ряда дескрипторов, представляемых в двоичной форме:

$$sb = (I_1, I_2 ... I_n, D_1, D_2 ... D_m, EN_1, EN_2 ... EN_k)$$

где $I_1, I_2 ... I_n$ – показатели исправности i-й подсистемы (i=1..n, n – количество подсистем, учитываемых при моделировании), принимающие значение 1, если подсистема исправна и 0 – в противном случае;

$D_1, D_2 ... D_m$ – показатели активности i-го механизма управления (i=1..m, m – количество управляющих механизмов, учитываемых при моделировании), принимающий значение 1, если механизм применяется; 0 – если механизм не задействован;

$EN_1, EN_2 ... EN_k$ – показатели, описывающие требования внешнего регулирования, принимающие значение 1, если объект соответствует некоторым нормам внешнего регулирования, 0 – в противном случае (индекс i определяет возможность рассматривать k различных аспектов внешнего регулирования).

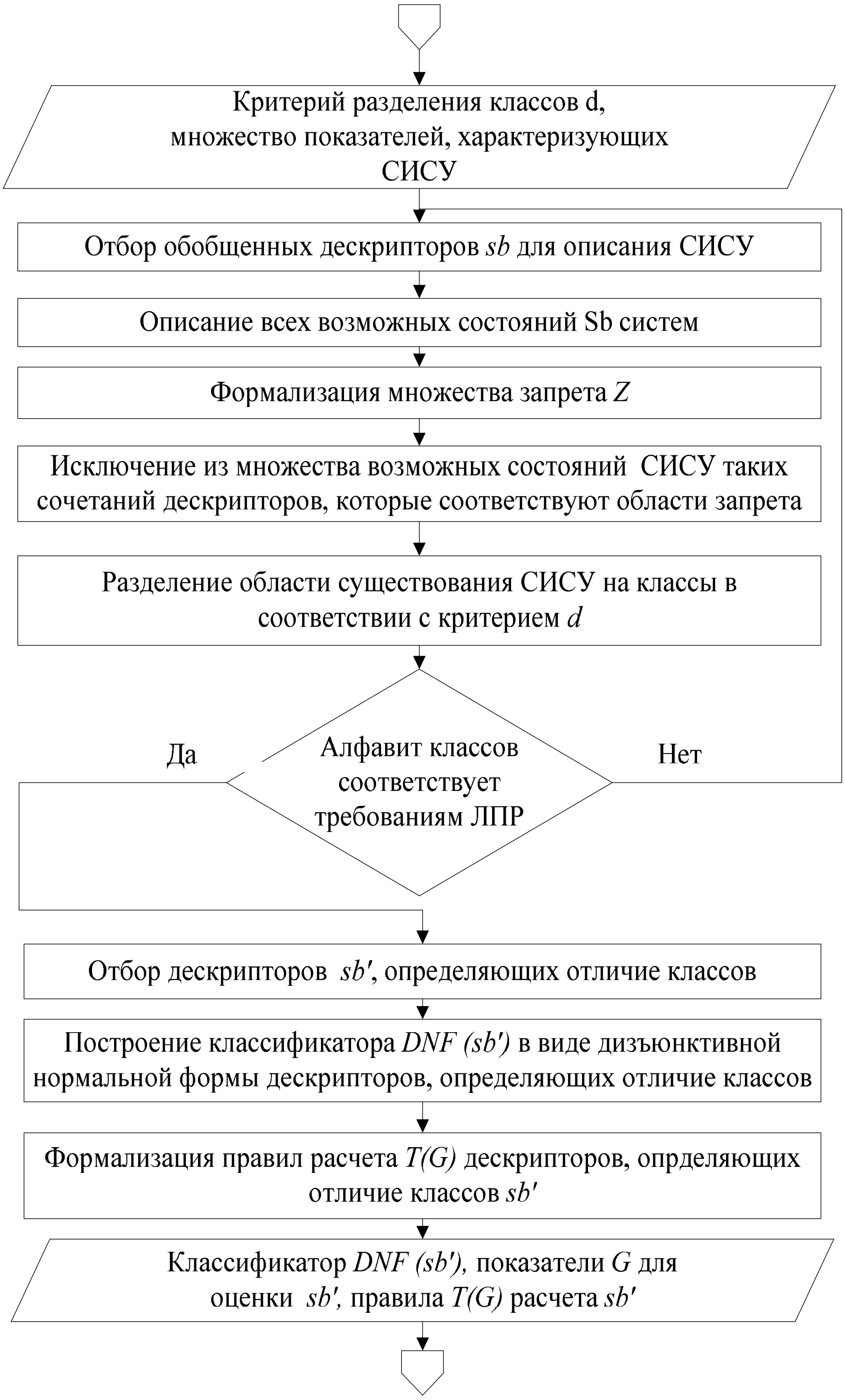


Рисунок 1 – Графическая схема алгоритма метода классификации без учителя на основе неколичественно заданного критерия разделения классов

На втором этапе выполняется описание всех возможных состояний систем, что равнозначно составлению всех возможных сочетаний $n+m+k$ двоичных признаков. В результате будет получено 2^{n+m+k} возможных сочетаний.

Описание области существования объекта классификации удобно представить в форме матрицы *Sb* размерностью $(n+m+k)\times(2^{n+m+k})$.

Решение практических задач приводит к составлению матриц большой размерности, т.к. адекватное описание большинства СИСУ требует десятков показателей. Количество возможных сочетаний достигает нескольких тысяч.

Однако значительная часть таких сочетаний невозможна в реальных условиях. С целью исключения противоречивых комбинаций предложенный метод предусматривает анализ взаимосвязей между показателями *I, D* и *EN*. Результатом является формализация множества запрета *Z* – правил, ограничивающих возможные сочетания значений дескрипторов [2].

Составление множества запрета в общем случае – сложная эвристическая задача, требующая детальных знаний о структуре и функционировании объекта классификации. С целью упрощения данного этапа метод предполагает последовательный поиск взаимосвязей различной природы и формализацию запретов следующих групп:

1) запреты физических противоречий Z^P, определяемые физическими законами существования и взаимодействия подсистем СИСУ;

2) запреты иерархии Z^H являются следствием иерархических связей между подсистемами СИСУ;

3) запреты рационального управления Z^{RM}, описывающие логику управления, нацеленного на повышение устойчивости, и применение для этого наиболее эффективных средств;

4) запреты внешнего регулирования Z^{EN}, т.е. ограничения возможностей функционирования и управления СИСУ со стороны внешних сил (законодательной системы, морально-этических норм, охраны труда и проч.).

Исключение из множества возможных состояний СИСУ такие сочетания дескрипторов, которые соответствуют области запрета *Z,* позволяет на один-два порядка сократить количество столбцов матрицы *Sb*.

Полученная в результате матрица *ZSb* содержит полные с точки зрения выбранной модели СИСУ описания состояний классифицируемых объектов. Сравнение этих описаний объединение их в группы, наиболее соответствующие требованиям критерия *d* позволяет сформировать ряд возможных алфавитов классов.

Если среди предложенных алфавитов классов нет соответствующего требованиям лица, принимающего решение (ЛПР), то построение классификатора в рамках существующего набора дескрипторов *sb* невозможно. Решение поставленной задачи в этом случае требует развития модели, используемой для описания состояния СИСУ, и повторения всех перечисленных выше этапов построения классификатора на основе другого множества показателей *sb*.

Если среди составленных алфавитов классов найден хотя бы один, удовлетворяющий требованиям лица, принимающего решение (ЛПР), то метод предполагает выявление среди дескрипторов *sb* таких показателей *sb'*, значения которых позволяют идентифицировать объекты различных классов.

Классификатор в результате представляет собой дизъюнктивную нормальную форму [2] от показателей *sb'*.

Классификатор в дизъюнктивной форме может использоваться как для бинарной классификации [5], так и для мультиклассовой. В этом случае число «слагаемых» должно быть на единицу меньше количества классов *w*. Тогда функция-классификатор сможет принимать значения от 0 до *w*-1, что достаточно для разделения *w* классов.

Завершение разработки классификатора требует формализации правил расчета обобщенных дескрипторов *sb'*, используемых при построении классификатора. Данный этап необходим, т.к. не всегда показатели *sb'*, позволяющие оптимально классифицировать СИСУ, поддаются непосредственному наблюдению. Определение значений многих величин требует дополнительных исследований или применения специаль-

ных средств измерения, выбор которых в значительной степени определяет качество дальнейшей классификации. Для измерения отдельных обобщенных дескрипторов необходимо рассчитать или измерять несколько показателей. Результатом является множество легко измеримых показателей *G*, а также правила *T(G)*, регламентирующие оценку показателей *sb'* по заданным значениям показателей *G*:

$$T(G) = sb'.$$

Двухэтапная процедура отбора показателей оправдана, так как применение на начальных этапах исследования обобщенных дескрипторов *sb'* предоставляет преимущества «прозрачности» при описании состояния СИСУ и при составлении множества запрета, а также позволяет получать более простые классификаторы. На завершающих этапах исследования необходимы легко измеримые показатели *G*, с целью исключения ошибок оценки входных параметров модели. Составление множеств *G* и *T* на последнем этапе разработки модели классификации значительно экономичнее, так как позволяет работать с показателями *sb',* количество которых в несколько раз меньше, чем *sb,* используемых на первых этапах.

Выводы

Таким образом, в результате проведенных исследований получен метод классификации, обладающий следующими особенностями:

1) не требует предварительного обучения на эталонной выборке, что существенно сокращает время разработки модели классификации и исключает влияние ошибки репрезентативности обучающей выборки;

2) позволяет разделять классифицируемое множество на два и более подмножеств. Количество возможных классов определяется количеством непротиворечивых сочетаний значений двоичных показателей, используемых для описания состояния СИСУ;

3) учитывает ограничения определяемые иерархией подсистем объекта классификации, физическими законами его функционирования, логикой рационального управления и нормами внешнего регулирования;

4) позволяет разделять объекты не произвольным случайным образом, а в соответствии с заданным критерием классификации;

5) позволяет использовать критерии разделения классов, заданные в произвольном виде, и при этом не требует упрощения задачи с целью приведения критерия к количественной форме;

6) позволяет получать классификатор в дизъюнктивной нормальной форме, что раскрывает широкие возможности автоматизированной и аппаратной реализации предложенного метода классификации.

В результате полученный метод предоставляет возможности решения задач, в равной степени недоступных для классических методов классификации с учителем и без учителя – задач с неполной информацией о составе классов, содержащих качественные критерии разделения классов.

Литература

1. Закревский, Д.А. Логика распознавания / Закревский Д.А. – Мн. : Наука и техника, 1988. – 118 с.
2. Чень Ч. Математическая логика и автоматическое доказательство теорем / Чень Ч. – М. : Наука, 1983. – 358с.

3. Головко В.А. Нейросетевые методы обучения и обработки информации в системах управления и прогнозирования : дис. ... д-ра тех. наук : 05.13.01 ; защ. : 14.01.03 ; утв. : 24.09.03 / Головко В.А. – Мн., 2002.
4. Корноушенко Е.В. Адаптивный алгоритм мультиклассовой классификации / Е.В. Корноушенко, А.А. Лобко // Международный конгресс по информатике: информационные системы и технологии = International Congress on Computer Science: Information Systems and Tecnologies : материалы международного научного конгресса, (Республика Беларусь, Минск, 31 окт. – 3 нояб. 2011 г.) : в 2 ч. Ч. 1 / редкол. : С.В. Абломейко (отв. Ред.) [и др.]. – Мн. : БГУ, 2011. – С.81-85.
5. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности : справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.Е. Енюков, Л.Д. Мешалкин; под ред. С.А. Айвазяна. – М. : Финансы и статистика, 1989. – 607 с.
6. Мандель И.Д. Кластерный анализ / Мандель И.Д. – М. : Финансы и статистика, 1988. – 176 с.

Literatura

1. Zakrevskij D.A. Logika raspoznavanija. Mn.: Nauka i tehnika. 1988. 118 s.
2. Chen' Ch. Matematicheskaja logika i avtomaticheskoe dokazatel'stvo teorem. M.: Nauka. 1983. 358 s.
3. Golovko V.A. Nejrosetevye metody obuchenija i obrabotki informacii v sistemah upravlenija i prognozirovanija: Dis. na soisk. uchen. step. d-ra tehn. nauk: 05.13.01: Zashhishhena 14.01.03: Utv. 24.09.03 Mn. 2002.
4. Kornoushenko E.V. International Congress on Computer Science: Information Systems and Tecnologies: materialy mezhdunarodnogo nauchnogo kongressa, Respublika Belarus', Minsk, 31 okt.-3 nojab. 2011 g. Mn.: BGU. 2011. S.81-85.
5. Ajvazjan S.A. Prikladnaja statistika: Klassifikacija i snizhenie razmernosti: Sparv. izd. M.: Finansy i statistika. 1989. 607 s.
6. Mandel' I. D. Klasternyj analiz. M.: Finansy i statistika. 1988. 176 s.

N.N. Masalitina

The Categorization Method without Learning Based on Qualitative Class Division Criteria

The results of solving the problem of categorization based on qualitative class division criteria are presented. The developed method allows to recognize more than two classes.

The method is based on the prohibition ensemble analysis. It allows to reduce a dimensionality of an object existence area and to unite objects in groups differ by a given criterion. As a result, the method allows to solve semiformalized problems of categorization when information about model inputs and outputs is not full.

The method does not require a learning stage. The whole class alphabetic is formed in a process of model synthesis and may be changed on any stage of development.

The method is orientated to the objects with a hierarchy structure (production systems, computing networks, organism and its functional subsystems in a treatment process and others).

The use of Boolean algebra principles gives wide possibility for automation of developing categorization models.

Статья поступила в редакцию 26.04.2012.